DOCUMENT RESUME

ED 170 365 TM 008 841

AUTHOR Roid, Gale; And Others
TITLE Item Writing for Domain-Based Tests of Prose Learning.
PUB DATE Apr 79
NOTE 36p.; Paper presented at the Annual Meeting of the American Educational Research Association (63rd, San Francisco, California, April 8-12, 1979)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Achievement Tests; *Complexity Level; *Criterion Referenced Tests; Elementary Education; *Multiple Choice Tests; *Prose; Reading Comprehension; *Test Construction; *Test Items; Test Reliability; Test Validity
IDENTIFIERS Distractors (Tests); Domain Referenced Tests

ABSTRACT

Differences among test item writers and among different rules for writing multiple choice items were investigated. Items testing comprehension of a prose passage were varied according to five factors: (1) information density of the passage; (2) item writer; (3) deletion of nouns, as opposed to adjectives, from the sentence in order to construct the item; (4) verbatim versus paraphrased use of keywords; and (5) use of an algorithm for choosing foil words (distractors) versus item writer's choice of foil words. Two dependent variables were examined: item difficulty, and variability of item difficulty across item writers for a particular technique. Test forms contained multiple choice items on a 1,000-word prose passage. One form was administered as a pretest, and a different form was administered as a posttest after instruction (reading the passages); subjects were 423 third to sixth grade students. Results indicated that passage density led to increased difficulty, particularly when freedom in foil construction was allowed; and that it produced bias in item writers in spite of other efforts to minimize bias. Variability of item difficulty across different writers was reduced when algorithms for writing linguistic-based items were used. (GDC)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED170365

ITEM WRITING FOR DOMAIN-BASED TESTS OF PROSE LEARNING

Gale Roid
Tom Haladyna
Joan Shaughnessy
Teaching Research Division
Oregon State System of Higher Education

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Gale Roid

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

TM008 841

This paper to be presented at the annual meeting of the American Educational Research Association, San Francisco, April 1979, Session 7.08.

FOREWORD

This research was supported by contract number MDA-903-77-C-0189 from the Defense Advanced Research Projects Agency. Contracting officer's technical representative for this project has been Dr. Pat-Anthony Federico of the Navy Personnel Research and Development Center, San Diego.

Acknowledgements are extended to Dr. John Bormuth of the University of Chicago and Dr. Jason Millman of Cornell University who provided consulting advice for this research. Appreciation is also extended to Dr. Harold F. O'Neil, Jr., of the Army Research Institute for his encouragement in the formative stages of this effort.

We also wish to acknowledge the assistance of Dr. Patrick Finn of State University of New York at Buffalo who supervised the writing of test items.

ABSTRACT

This study examined several methods of writing test questions that may solve the problems of optimizing the match between teaching and testing and controlling item-writer differences. Specifically, rules for writing test items from prose instructional material were given to four item writers who created multiple-choice test items. Differences among item writers and among a variety of item writing rules were examined. Tests were given to 423 students before and after reading two prose passages. Methods of transforming sentences from prose material into test questions were found to control the typical variance of item difficulty that is observed between item writers. The information density of the prose passage, the method of writing foils (wrong answer alternatives), the part of speech of the keywords in the sentences transformed into questions, and verbatim vs. paraphrase use of sentences had important influence on the statistical characteristics of items.

INTRODUCTION

Problem

One of the most prevalent concerns of students in college and training is the match between teaching and tests. When there is close correspondence between the intent of the instructor and the test, student learning is likely to be very effective (Duchastel & Merrill, 1973; Hartley & Davies, 1976; Seddon, 1978). Concern for the match between teaching and tests has prompted several measurement theorists, such as Anderson (1972) and Bormuth (1970), to propose a technology of test item writing in which the generation of test items is operationally defined as a transformation of elements in instruction. If the intent behind an instructional program is developed into a definition of a domain or universe of content, then a pool of test items can be specified to match that domain or universe. A domain-based test (Millman, 1974; Hively, 1974) is, then, a random sample of items from that domain. Hambleton, Swaminathan, Algina and Coulson (1978), among others, have chosen to refer to domain-based tests as the purest form of criterion-referenced tests and prefer the latter term. The development of domain specifications for improving criterion-referenced tests have been a topic of great concern for measurement specialists (Popham, 1978).

Background

One area of research in domain-based testing is when the content to be learned is based on prose (Bormuth, 1970). Bormuth described a technology for operationally defining a set of rules which can be given to an item writer who transforms elements of a prose passage into test questions. This linguistic-based theory of item writing has subsequently undergone further research and development (Finn, 1975; Roid & Finn, Note 1; Roid, Haladyna &

Finn, Note 2). One branch of linguistic-based item writing is the transformation of sentences from prose instructional materials into multiple-choice test questions. The use of sentences as a basic element of instruction is based on the logic that these are countable units of instruction which define a domain of content used in a course of instruction. Also, as Bormuth (1970) has described, by transforming sentences from instructional materials, it is possible for the difficulty of tests to be matched exactly to the readability level of the instructional materials and, therefore, creating the purest form of a criterion-referenced test. The sentence-based technology of item writing is, of course, most appropriate for the recall or comprehension level of cognitive learning. There are many objectives of instructional programs that require, at least as enabling objectives, the learning of terminology and basic meanings from prose segments of instruction. For example, in dentistry a basic course is dental anatomy which requires the recall and comprehension of names, locations, contours and morphology of each of the adult and primary teeth. The authors have found the methods of Bormuth highly useful in producing domain-based tests for instructional materials in such content areas.

Purpose

The purpose of the present study was to further refine and empirically study a method for objectively generating multiple-choice questions by transforming sentences from prose instructional materials. Recent research by the authors, Roid, Haladyna and Finn (Note 2), have revealed that Bormuth's theory and methods for producing test items are feasible, but more research is needed concerning refinements of this approach in contrast to more familiar and less automated item-writing approaches. The present study was

planned to investigate some of the characteristics of items under the influence of five independent variables:

1. The information density of prose passages was studied to determine if the amount of information in a passage affects the difficulty of test items.

2. Different item writers were compared to test the hypothesis that specific sentence transformation rules reduced the variability of item difficulties across item writers.

3. Nouns and adjectives were compared as answer words to be deleted in transforming sentences into test questions.

4. Verbatim use of the sentence from the passage was compared to paraphrasing of keywords in the sentence.

5. An algorithm for choosing foil words from the passage was compared to item writers who selected the wording of foils for parallel forms of the items.

There are several important reasons for studying these five independent variables. First, if domain-based tests of prose learning are to result in a better match between the difficulty of a passage and the test created for it, then the information density or difficulty of the passage should be varied to study this effect. Secondly, if domain-based methods of test construction are to improve the technology of item writing, then they should reduce the variability between item writers so that when tests are used in a mastery learning system, the difficulty and mastery level (cut-off score) can be predictably determined. Third, the technology appears to work best with nouns and adjectives as opposed to verbs and adverbs, and this needs to be further confirmed by research. Fourth, Anderson (1972) argued convincingly for the importance of paraphrasing in the writing of test items for

comprehension. By systematically varying this element in the present study, the distinction between recall and comprehension level questioning should be explicated. Fifth, the production of quality foils or wrong answers for multiple-choice questions is a time-consuming and challenging part of item writing. If automated or algorithmic methods of foil construction are feasible and can create items of acceptable quality, the technology of item writing would move forward in providing practical advantages to test developers.

METHOD

Instructional Materials

Two prose passages, each approximately 1,000 words in length, were selected from a popular children's wildlife magazine. The overall reading levels (Dale & Chall, 1948) for passage one (score 5.359) and for passage two (score 5.261) were compared and the difference between them was non-significant.

In order to assess the information content of the passages, copies of each passage were distributed to 17 intermediate grade teachers in four elementary schools. Each teacher was asked to underline all sentences in both prose passages which contained information they believed to be important for their students to remember. The teachers underlined an average of 42.42% of the sentences in passage one, but only an average of 33.63% of the sentences in passage two. This difference between passages was statistically significant (correlated $t = 6.11$, $df = 16$, $p < .001$). Passage one was, therefore, determined to contain more of the information which teachers believed to be important for student learning. Since the passages were of equivalent length, the teachers were reporting that passage one contained a greater density of information than passage two.

Item and Test Development

All nouns and adjectives which appeared only once in a single passage and which had a low standard frequency index, that is, are relatively rare in American textbooks (Carroll, Davies & Richman, 1971), were listed. Four adjectives and four nouns from each passage (a total of eight adjectives and eight nouns) were randomly selected from this "rare singleton" list and the 16 sentences in which these words appeared then were the basis for test items. The selected sentences were transformed using the method of Finn

(Note 3) into multiple-choice items by four item writers. This transformation method is described in detail in Roid and Finn (Note 1). All item writers followed identical rules for item transformation and for sentence paraphrasing. Each item writer produced two questions from each of the 16 selected sentences, one question with a verbatim sentence-transformation and the other paraphrased. Paraphrasing was accomplished by the item writer paraphrasing one keyword in the question stem and paraphrasing one answer word among the four multiple-choice words.

Two different foil methods were used. For the first foil method, each writer personally selected foils or alternatives for each question. In the second foil method, an algorithmic system for generating the three foils for each test question was developed. In the case of noun-based items, foils for each item were selected from a list of all rare singleton nouns with the same semantic classification (Frederiksen, 1975) as the answer noun. For adjective-based items, all rare singleton adjectives were classified using semantic differential techniques (Nunnally, 1967, pp. 536-538) and foil adjectives were randomly selected as possible alternatives from the list of adjectives in the same classification as the adjective answer.

This process resulted in 256 items, consisting of 16 selected sentences both transformed and paraphrased by four item writers using two foil methods. From these 256 items, 16 test forms, each 16 questions in length, were formed. Each test form included an equal number of noun-based and adjective-based items from each passage. Half of the questions were paraphrased and the other half were verbatim sentence transformations. Also, each test contained an equal number of items having writer's-choice foils versus algorithmic foils.

Procedure

The 16 test forms were administered to 423 third, fourth, fifth and sixth grade students in four schools. Each student was randomly assigned one of the 16 test forms as a pretest prior to reading either passage. Then after the class read the passages together and the teacher reviewed the content of the passages, the posttests were administered at an interval of two to five class days after the pretest. For the posttest, students were randomly assigned one of the 15 forms other than their pretest form.

Analysis

Two types of dependent variables were employed in this study. The first was the item difficulty (the percentage of correct responses for all examinees), and the second was the variability of the item difficulties across item writers for a particular technique or combination.

To answer the research questions addressed earlier in this paper, item difficulties were subjected to an analysis of variance based on each of the two dependent measures. Therefore, a 2 x 4 x 2 x 2 x 2 x 2 analysis of variance was employed. The factors were high vs. low information density of the prose passage (I), four item writers (W), noun vs. adjective part of speech (P), paraphrase or verbatim stem contraction technique (S), algorithm or writer's choice foil construction (F), and the repeated measure (R)--pretest and posttest.

For the first dependent variable, the statistically significant interaction of any factor with the repeated measure is an indicator of what could be called "instructional sensitivity." Assuming instruction is effective (as judged by the means of pretest and posttest scores), a significant interaction of any independent variable with the repeated measure (R) suggests that one type of test item is more sensitive to the effects of

instruction, i.e., it has instructional sensitivity. Another measure of instructional sensitivity is the difference between pretest and posttest item difficulties (Cox & Vargas, Note 4; Roid & Haladyna, 1977, p. 39).

For the second dependent variable, the analysis of variance was performed using the natural logarithms of the variances of item difficulties, as recommended by Scheffé (1959, p. 18). This procedure parallels the conventional analysis of variance, as performed on the first dependent variable, with the major exception being that the log variance of item difficulties are entries in the cells of the factorial design.

With any analysis of variance where sample sizes are large, there is a possibility of detecting statistical significance when the differences among groups is rather small. Although the null hypothesis is rejected, the practical significance of such findings is minimal. Therefore, to protect against such an eventuality, a test of practical significance was applied to all statistically significant results, according to the advice of Tallmadge and Horst (1975). To be practically significant, the difference between the mean of any level of a main effect or any level of a simple effect in an interaction and the grand mean should exceed one-third of a standard deviation.

RESULTS AND DISCUSSION

Item Difficulty

All statistically significant results for the first dependent variable are reported in Table 1. Six main effects and a total of 56 interactions were possible. The five-way interactions were considered too complex for analysis and, so, the sums of squares and respective degrees of freedom for these effects were pooled with the residual variance.

Insert Table 1 about here

Main effects for all six independent variables were statistically significant ($p < .05$), but only the repeated measure (R) and the passage density (I) were practically significant. Instruction was not highly effective, as judged by the pretest and posttest means of 46.4% and 65.4%, respectively. Items from the high density passage were more difficult overall, 48.2% as compared to low density passages, 63.8%. Thus, passage density significantly affected test performance.

Of the 50 interactions tested, only ten were statistically significant, where we would expect about two or three by chance alone. Of these ten significant findings, eight were practically significant and become the objects of further discussion.

In Table 2, the cell and marginal means and standard deviations are presented for three practically significant interactions: information density by foil method (I x F), part of speech by foil (P x F) and the three-way interaction, I x P x F. The I x F interaction was apparently due to the fact that foils created by item writers for the high density passage were particularly difficult (a mean of 43.8%) while foils created by algorithm were easier (60.4%).

Insert Table 2 about here

The P x F interaction was attributed to the fact that noun-based item stems, for which item writers produced the foils, were the most difficult (46.1%). All other conditions in the cells produced by the interaction of P and F were not particularly high or low.

The interaction among factors I, P, and F was practically significant principally due to the fact that items produced from the higher density passage, where a noun stem technique and either foil technique were used, were more difficult. Means were 41.9 for writer's choice and 45.2 for algorithmic foils, as shown in Table 2. A noun-based or adjective-based item stem coupled with an algorithmic foil technique used on the low information density passage proved to be easier than other conditions, where means were 69.6 and 68.9, respectively.

The three practically significant interactions (I x F, P x F, and I x P x F) provide support for the notion that passage density is a powerful and seemingly complex factor in determining item difficulty. Given that noun-based items are harder than adjective-based items, the combination of noun-based items for a dense passage resulted in very difficult items, where, surprisingly, foil technique does not seem to be a strong influence. What is particularly interesting are the large-magnitude differences between items involving algorithm-produced foils for high density and low density passages. Across noun and across adjective stem conditions, these differences were 24.7% (69.9%-45.2%) and 14.9% (68.9%-54.0%), respectively. The methods used to create items play a major role in determining difficulty in a seemingly complex yet understandable way. That is, highly dense passages

lead to more difficult questions. When coupled with noun-based techniques, the two factors jointly influence item difficulty.

Interactions among the factors I, W (item writers) and R are shown in Table 3. For the more dense passage, unusually low pretest and posttest means were exhibited for items produced by item writers 1 and 3. For the less dense passage, item writers 2 and 4 showed significantly higher means. It is clear from this interaction that item-writer bias is strongly affected by the density of the passage. While the overall difference among item writers was not practically significant as a main effect, this one interaction again points to the pervasive influence of passage density. Interestingly, interactions of item writers with variables such as stem type or foil were not practically significant, thus lending evidence to the proposition that these procedures reduce the bias that is prevalent when item writers are given more freedom in stem and foil construction, as witnessed in the study by Roid and Haladyna (1978).

Insert Table 3 about here

Table 4 shows the cell and marginal means for the practically significant interactions of R x P x F and R x P x S x F. Only one condition for the R x P x F effect produced practically significant results: where the stem was noun-based and where the item writer wrote the foil. For both pretest and posttest conditions, as shown in the bottom, marginal means in Table 4, noun-based writer's choice items were relatively difficult, 34.0 and 58.1, respectively. In terms of instructional sensitivity, however, the pretest-posttest shift was 24.1% as contrasted with the three other conditions where the shifts were 14.1%, 19.3% and 18.3%. Although items based on nouns, with

writer's-choice foils, were more difficult, they were more sensitive to the effects of instruction. From this standpoint, these items may be viewed as more appropriate in terms of reflecting the shift in performance due to instruction.

Insert Table 4 about here

The R x P x S x F interaction revealed that the type of stem construction, paraphrase or verbatim, does interact with noun-based or adjective-based deletions in key sentences as well as foil approach. Nine of 12 means were practically significant, and the patterns of these results were quite definitive. Verbatim item stems were relatively easy, with the exception of the condition where the noun-based stem technique and the writer's-choice foil technique were used. Paraphrase items were consistently more difficult regardless of pretest or posttest means. In terms of instructional sensitivity, noun-based, verbatim item stems where foils were produced by writers were very sensitive as were verbatim items where algorithms were used. This curious state of affairs may be explained by the general tendency of verbatim items to be both easier and more sensitive to instruction regardless of whether the keyword deletion involves nouns or adjectives. It is clear that the paraphrase items were both difficult and less sensitive to instruction. An inevitable dilemma remains: Is this lack of sensitivity to instruction a product of a poor learning experience on the part of the students with respect to comprehension level learning, or the tendency of these items not to be indicators of what was taught? This seems to be a problem which faces anyone who must develop and evaluate instruction, and can be resolved only by a further investigation of the instructional program or test items to determine the true sources of the effect.

Table 5 provides the cell and marginal means for the interactions of R x S and R x I x S.

Insert Table 5 about here

The R x S interaction is an important one as it provides a key to the effectiveness of a set of items as a function of instructional sensitivity. Items which are sensitive to instruction have been viewed as more valid indicators of student learning (Glaser, 1971, p. 12). The R x S interaction clearly indicates that items written at the verbatim level, as contrasted to items written using the paraphrase technique, which appear to measure learning at the comprehension level, were more sensitive to instruction in this experiment. Paraphrase items showed an instructional sensitivity of 13.3% as contrasted to verbatim items which showed a gain from pretest to posttest of 24.6%. As reported earlier, this difference may be interpreted in several ways: (a) instruction was heavily oriented toward memorizing, (b) verbatim items are inherently easier while comprehension items are inherently harder, (c) verbatim items are sensitive to what students learned while comprehension items are not, or (d) paraphrasing introduced new vocabulary with which the students were unfamiliar, resulting in more difficult, non-instructionally relevant items. While it is not clear what factors underlie this difference or which interpretation is most truthful, it is clear that the differences exist as a function of cognitive level and that the higher level appears more difficult and less sensitive to instruction.

The last interaction to be discussed, R x I x S, reveals the relation of passage density to cognitive level. Passage density has been shown to

powerfully interact with a number of other variables in producing results which suggest a complex force which affects both item difficulty and instructional sensitivity. With the R x I x S interaction, five of the eight means significantly departed from the respective pretest and posttest grand means. With the more dense passage and the paraphrase condition, as expected, pretest and posttest means were quite low as was instructional sensitivity. With the less dense passage and verbatim stem construction, pretest and posttest means were quite high. In both of these extreme conditions, instructional sensitivity was rather low, 11.5% in one instance and median in the other 18.8%. With the other two conditions, verbatim-more dense and paraphrase-less dense, item difficulties were more in the normal range with a relatively high instructional sensitivity, 30.5% for the former, and a medium index for the latter of 15.1%. Thus, it would seem that verbatim-level questioning for more dense passages are highly sensitive to instruction, while other conditions are not so sensitive. Use of the paraphrase stem with the more dense passage resulted in significantly more difficult test items which are highly insensitive to instruction. One is compelled to question whether such items are actually less valid or whether instruction needs to emphasize higher level learning instead of the verbatim material that is possibly easier to learn.

Item Variability

A separate analysis of variance dealing with the variances of item difficulties was conducted and revealed that only the independent variable of foil technique produced sizable differences in these variances. The main effect for foil technique was statistically significant at $F = 9.66$, for $df = 1,48$ with $p < .003$. The algorithmic foil technique produced item

difficulties that varied less across item writers than items where foils were produced by the item writer. This result confirms the hypothesis that the algorithm for producing foils provides a control on item-writer bias.

From analyses discussed above, it is clear that items with algorithm-based foils were neither significantly weaker in terms of instructional sensitivity, nor more or less difficult, than other items in this study. Further, the algorithm-based items were not influenced greatly by other factors, with the exception of passage density.

CONCLUSIONS

A number of techniques were studied for algorithmically constructing multiple-choice test items to assess learning from prose material. Most of these techniques were originally based on Bormuth's theory of achievement testing (Bormuth, 1970). These linguistically-based procedures were proposed by Bormuth to reduce item-writer bias which is prevalent when the item writer has freedom to write stems and foils for multiple-choice items. Working with Finn's extension of Bormuth's work (Finn, 1975), refinements were made in this methodology (Roid & Finn, Note 1), and the present study serves to delineate some of these refinements as well as compare alternative strategies in stem and foil construction across varying conditions. The findings for the first dependent variable of the study, item difficulty, will be organized around the major independent variables of the study in order of their importance: (a) passage density, (b) foil construction, (c) item writer, (d) stem technique, and (e) the part of speech of deleted keywords.

Effects on Item Difficulty

Passage density. An important finding of this study was that passage density is a powerful and complex factor which affects test performance as well as learning. While it is virtually impossible to divorce the characteristics of the test from the student learning that actually occurs, it is clear that a more dense passage leads to lower test performance and less gain instructionally as judged by pretest and posttest shifts in means.

In fact, passage density was involved in four of eight practically significant interactions. Although foil technique was involved in five of these eight interactions, seldom did foil technique interact as powerfully or systematically as passage density.

Passage density appears to be a major determinant of item difficulty. Passage density is related in a complex way to the factors of (a) part of speech deleted, (b) stem technique (verbatim or paraphrase), (c) item writer, and (d) foil technique. For example, noun-based items for a more dense passage (with either foil technique) produced unusually difficult items, while in a less dense passage, the algorithm foil technique produced easier items regardless of the part of speech deleted. When studied in relation to the cognitive level of stem construction, passage density again significantly interacted. Paraphrase items which measure comprehension were more difficult for the more dense passage, while verbatim items which measure recall were easier for the less dense passage. Instructional sensitivity was highest in the condition where verbatim items were used with the more dense passage.

Foil construction. While passage density was a powerful independent variable, the type of foil construction, as another independent variable, was involved in more interactions. In five of eight practically-significant interactions, foil construction was a factor.

An important interaction was that passage density and foil construction joined to make item difficulty extremely high or low. More dense passages where item writers had freedom in foil construction produced hard items, while in less dense passages where no freedom was permitted and algorithms were used, items were easier.

Item writer. Item-writer bias, overall, can be significantly reduced through the employment of many of these algorithms or other rules of item writing. However, the item writer factor was found to interact with passage density. It is clear that the passage density factor can produce bias in item writers despite the effects of other variables to reduce this bias.

Future efforts need to concentrate on controlling passage density as a determinant of item difficulty and item-writer bias.

Stem technique. Items produced using the verbatim stem technique are designed to measure recall, while items written using the paraphrase technique are designed to measure comprehension. In three practically significant interactions, this factor was involved. Generally, verbatim items are easier across all conditions. However, stem technique did significantly interact with the repeated measure. Larger gains were observed from pretest to posttest for verbatim items as compared to paraphrase items. And, as reported earlier, verbatim items showed a larger pre-to-post shift for the more dense passage.

Part of speech of deleted keywords. Either a noun or an adjective (or its phrase) were deleted in transforming a sentence into a question. Overall, noun-based items were harder than adjective-based items, although this difference was not practically significant. Part of speech did interact with passage density, foil technique, and the repeated measure. The patterns of these interactions were generally unclear. Upon reflection, the choice of noun or adjective would be expected to have little to do with the characteristics of items, but, in fact, it is the interaction of the noun-adjective factor with other factors, such as passage density or foil technique, that affect item characteristics in a complex way.

Summary. In conclusion, these results point to the importance of analyzing the characteristics of a prose passage when using linguistic-based item writing methods. Passage information-density appears to have much to do with the difficulty of items. If item difficulties were allowed to vary in unpredictable ways across passages in systematic instruction, where passing standards are used, students' successes may be more of a function of which set of items are used rather than what the students have learned.

Some of the evidence in this study emphasizes the need for attention to the role that item writers play. In some instances, letting item writers choose foils appears to increase instructional sensitivity, but if the item writer additionally uses paraphrasing, instructional sensitivity can be dramatically reduced. One possible explanation for this effect of paraphrasing is that new words of higher reading difficulty that are unfamiliar to the student may be introduced into test items.

Verbatim transformations of sentences produced more instructionally sensitive items, but this may have happened as a function of the method by which foils were produced. The combination of verbatim item transformation of noun-based sentences coupled with writer's-choice foils was the most superior technique identified in this study.

Effects on the Variance of Item Difficulties Across Item Writers

A very important finding of this study is that control of item difficulty is possible and a measurable reduction in item-writer bias can be obtained via item writing rules such as those implemented in this experiment.

Variability of item difficulties across item writers was found to be significantly less when an algorithm was used to create foils as compared to each writer selecting the wording of foils. This shows that the idiosyncracies of wording in foils created by item writers are a powerful source of bias. If the goal of a team of item writers is to write items of similar difficulty for use in criterion-referenced tests, then, clearly, attention should be paid to foil-writing methods. The present study indicates that operationally-defined rules for creating foils provide a quality control mechanism that creates items of more homogeneous difficulty. Control of item difficulty is crucial for tests to be used for pass/fail decisions with

a fixed score criterion because item difficulty directly affects the distribution of scores on such a test. Field testing of the items may be used, after the fact, to assess item difficulties for purposes of adjusting them, but attention to item-writer differences in the first stages of test development could be less expensive and create a test of more predictable statistical qualities.

Another benefit of operational definitions of foil construction would be that the resulting tests are then precisely described to other researchers and the public. Clearly, an admirable goal of any test development is to have available a precise definition of what is being measured and counted by the test (Anderson, 1972, p. 145). The algorithms of the present study are one alternative in providing such a precise definition.

Future studies could explore refinements of these methods for item writing as well as isolating the role that passage density plays in determining item difficulty. Inevitably, the approach to testing studied here must be integrated with a theory of instruction that takes into account the levels of mental processes involved. In a related way, the work of Scandura (1978) and of Landa (1974), who deal with structural aspects of learning, seems important.

More importantly, there is a growing recognition of the importance of the relationship between instruction and achievement testing that is the hallmark of systematic instruction and criterion-referenced testing. Systematic algorithmic test item writing should become a necessary step in CR test development, and this study documents one aspect of this emerging technology. Algorithmic approaches to test item writing provide an alternative to the problems of domain specifications which precede the creation of criterion-referenced tests.

REFERENCES

Anderson, R. C. How to construct achievement tests to assess comprehension. Review of Educational Research, 1972, 42, 145-170.

Bormuth, J. R. On the theory of achievement test items. Chicago: University of Chicago Press, 1970.

Carroll, J. B., Davies, P., & Richman, B. Word frequency book. Boston: Houghton-Mifflin, 1971.

Dale, E., & Chall, J. S. A formula for predicting readability. Educational Research Bulletin, 1948, 27, 11-28.

Duchastel, P. C., & Merrill, P. F. The effect of behavioral objectives on learning: A review of empirical studies. Review of Educational Research, 1973, 43, 53-69.

Finn, P. J. A question writing algorithm. Journal of Reading Behavior, 1975, 4, 341-367.

Frederiksen, C. H. Representing logical and semantic structure of knowledge acquired from discourse. Cognitive Psychology, 1975, 7, 371-458.

Glaser, R. Instructional technology and the measurement of learning outcomes: Some questions. In W. J. Popham (Ed.), Criterion-referenced measurement: An introduction. Englewood Cliffs, N.J.: Educational Technology Publications, 1971.

Hambleton, R. K., Swaminathan, H., Algina, J., & Coulson, D. B. Criterion-referenced testing and measurement: A review of technical issues and developments. Review of Educational Research, 1978, 48, 1-47.

Hartley, J., & Davies, I. K. Preinstructional strategies: The role of pretests, behavioral objectives, overviews and advance organizers. Review of Educational Research, 1976, 46, 239-265.

Hively, W. Introduction to domain-referenced testing. Educational Technology, 1974, 14, 5-10.

Landa, L. N. [Algorithmization in learning and instruction] (F. F. Kopstein, Ed., V. Bennett, Trans. from Russian). Englewood Cliffs, N.J.: Educational Technology Publications, 1974.

Millman, J. Criterion-referenced measurement. In W. J. Popham (Ed.), Evaluation in education: Current applications. Berkeley, California: McCutchan Publishing Company, 1974.

Nunnally, J. Psychometric theory. New York: McGraw-Hill, 1967.

Popham. W. J. Criterion-referenced measurement. Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1978.

Roid, G., & Haladyna, T. Measurement problems in the formative evaluation of instructional systems. Improving Human Performance Quarterly, 1977, 6(1), 30-44.

Roid, G. H., & Haladyna, T. M. A comparison of objective-based and modified-Bormuth item writing techniques. Educational and Psychological Measurement, 1978, 35, 19-28.

Scandura, J. M. A deterministic theory of teaching and learning. In H. Spaga & W. F. Kempf (Eds.), Formalized theories of learning and thinking. Bern: Huber, 1976.

Scheffé, H. The analysis of variance. New York: Wiley, 1959.

Seddon, G. M. The properties of Bloom's taxonomy of educational objectives for the cognitive domain. Review of Educational Research, 1978, 48, 303-323.

Tallmadge, G. K., & Horst, D. P. A procedural guide for achievement gains in educational evaluation. Mountain View, California: RMC Research Corporation, 1975.

REFERENCE NOTES

1. Roid, G. H., & Finn, P. J. Algorithms for developing test questions from sentences in instructional materials (NPRDC Tech. Rep. 78-23). San Diego: Navy Personnel Research and Development Center, June 1978.
2. Roid, G., Haladyna, T., & Finn, P. A comparison of several multiple-choice, linguistic-based item writing algorithms. Manuscript submitted for publication, Monmouth, OR: Teaching Research, 1978.
3. Finn, P. J. Generating domain-referenced, multiple-choice test items from prose passages. Paper presented at the meetings of the American Educational Research Association, Toronto, March 1978.
4. Cox, R. C., & Vargas, J. A comparison of item selection techniques for norm-referenced and criterion-referenced tests. Paper presented at the annual meeting of the American Educational Research Association, 1966.

Table 1

Summary of Statistically and Practically Significant[1]

Results of the Analysis of Variance Using

Item Difficulty as the Dependent Variable

Source	df	F	p	Practically significant
Passage density (I)	1, 208	73.24	< .001	yes
Writer (W)	3, 208	8.46	< .001	no
Part of Speech (P)	1, 208	20.74	< .001	no
Stem Technique (S)	1, 208	47.69	< .001	no
Foil Technique (F)	1, 208	16.08	< .001	no
I x F	1, 208	6.13	.014	yes
W x F	1, 208	4.20	.007	no
P x F	1, 208	5.75	.017	yes
I x P x F	1, 208	4.22	.041	yes
Repeated Measures (R)	1, 208	329.66	< .001	yes
R x S	1, 208	31.41	< .001	yes
R x F	1, 208	7.41	< .001	no
R x I x W	3, 208	4.53	.004	yes
R x I x S	1, 208	13.53	< .001	yes
R x P x F	1, 208	4.93	.033	yes
R x P x S x F	1, 208	4.90	.028	yes

[1] Practical significance was determine in the following way:

$\bar{X} - \bar{\bar{X}} = \frac{sd}{3}$ where $\bar{X}$ is the mean of any level of a main effect or any level of a simple effect in an interaction and $\bar{\bar{X}}$ is the grand mean of that main or simple effect.

Table 2

Cell and Marginal Means for Practically Significant Interactions of I x F, P x F and I x P x F

		Writer's Choice Foil (F1)			Algorithm Foil (F2)		
		More Inform. Density (I1)	Less Inform. Density (I2)	Total (I1 + I2)	More Inform. Density (I1)	Less Inform. Density (I2)	Total (I1 + I2)
Noun-based Stem (P1)	mean	41.9*	50.3	46.1*	45.2*	69.9*	57.2
	s.d.	20.4	25.9	23.6	19.0	18.4	22.4
	n	64	64	128	64	64	128
Adjective-based Stem (P2)	mean	51.9	65.3	58.6	54.0	68.9*	61.5
	s.d.	23.3	21.8	23.5	20.5	18.1	20.7
	n	64	64	128	64	64	128
Total	mean	46.9*	57.8	52.4	49.6	69.4*	59.5
	s.d.	22.4	25.0	24.3	20.1	18.2	21.6
	n	128	128	256	128	128	256

* An asterisk marks the means that are more than one-third a standard deviation different from the grand mean for the main or simple effect.

Table 3

Cell and Marginal Means[1] and Standard Deviations for Practically Significant Effect of Passage Information Density (I) and the Interaction of I, Writers (W) and the Repeated Measures (R)

		Writer 1 (W1)		Writer 2 (W2)		Writer 3 (W3)		Writer 4 (W4)		Total for Factor
		Pre	Post	Pre	Post	Pre	Post	Pre	Post	I Main Effect
More Information Density (I1)	mean	36.47*	57.06*	48.16	60.41	30.56*	52.19*	35.78*	65.22	48.2*
	s.d.	19.06	20.77	17.30	16.43	15.70	20.83	15.02	16.79	21.3
	n	32	32	32	32	32	32	32	32	256
Less Information Density (I2)	mean	53.19	69.25	59.28*	79.06*	50.12	65.84	57.97*	74.16*	63.6*
	s.d.	23.10	25.75	20.44	12.05	22.87	23.22	18.77	18.03	22.4
	n	32	32	32	32	32	32	32	32	256

[1] Those means marked with an asterisk were more than one-third of a standard deviation from the respective grand mean for the main or simple effect. Such deviations were grounds for practical significance.

Table 4

Cell and Marginal Means[1] and Standard Deviations for the Interactions of R x P x F and R x P x S x F

		Noun (P1)				Adjective (P2)			
		Writer's Choice (F1)		Algorithm (F2)		Writer's Choice (F1)		Algorithm (F2)	
		pretest (R1)	posttest (R2)	pretest (R1)	posttest (R2)	pretest (R1)	posttest (R2)	pretest (R1)	posttest (R2)
Paraphrase (S1)	mean	34.22*	48.06*	45.03	54.91*	43.34	60.03	49.69	62.44
	s.d.	16.57	20.53	17.77	23.57	20.47	23.61	22.80	18.61
	n	32	32	32	32	32	32	32	32
Verbatim (S2)	mean	33.84*	68.16	56.06*	74.34*	54.44*	76.44*	54.91*	78.81*
	s.d.	18.46	20.89	22.70	14.18	23.33	12.11	17.54	18.65
	n	32	32	32	32	32	32	32	32
Total	mean	34.03*	58.11*	50.54	64.62	48.89	68.24	52.30	70.62
	s.d.	17.40	22.91	20.97	21.64	22.48	20.37	20.35	20.24
	n	64	64	64	64	64	64	64	64

[1] Those means marked with an asterisk were more than one-third of a standard deviation from the respective grand mean for the main or simple effect. Such deviations were grounds for practical significance.

Table 5

Cell and Marginal Means[1] and Standard Deviations for Practically Significant Effects of the Repeated Measures (R) and the Interactions of R and the Stem Generating Technique (S) As Well As the Interactions of R and Passage Information Density (I) and S

		Paraphrase (S1)		Verbatim (S2)		Total	
		Pretest (R1)	Posttest (R2)	Pretest (R1)	Posttest (R2)	Pretest (R1)	Posttest (R2)
More Dense Passage (I1)	mean	36.80*	48.27*	38.69*	69.17	37.74	58.72
	s.d.	17.85	18.80	17.95	12.97	17.85	19.21
	n	64	64	64	64	128	128
Less Dense Passage (I2)	mean	49.34	64.45	60.94*	79.70*	55.14	72.08
	s.d.	20.45	22.38	20.94	15.16	21.42	20.52
	n	64	64	64	64	128	128
Total	mean	43.07	56.36	49.81	74.44*	46.44*	65.40*
	s.d.	20.12	22.13	22.43	15.01	21.53	20.93
	n	128	128	128	128	256	256

[1] Those means marked with an asterisk were more than one-third of a standard deviation from the respective grand mean for the main or simple effect. Such deviations were grounds for practical significance.